基于深度学习的中文专利自动分类方法研究*

■ 吕璐成[1,2] 韩涛[1,2] 周健[3] 赵亚娟[1,2]

[1] 中国科学院文献情报中心 北京 100190 [2] 中国科学院大学经济与管理学院图书情报与档案管理系 北京 100190

[3] 中国科学院计算技术研究所 北京 100190

摘 要：［目的/意义］面向当前国内专利审查和专利情报分析工作中对于海量专利分类的客观需求，设计了7种基于深度学习的专利自动分类方法，对比各种方法的分类效果，从而助力专利分类效率和效果的提升。［方法/过程］针对传统机器学习方法存在的缺陷，基于Word2Vec、CNN、RNN、Attention机制等深度学习技术，考虑专利文本语序特征、上下文特征以及分类关键特征，设计Word2Vec + TextCNN、Word2Vec + GRU、Word2Vec + BiGRU、Word2Vec + BiGRU + TextCNN等7种深度学习模型，以中国专利为例，选取IPC主分类号的"部"作为分类依据，对比这7种模型与3种传统分类模型在中文专利分类任务中的效果。［结果/结论］实证研究效果显示，采用考虑语序特征、上下文特征及强化关键特征的深度学习方法进行中文专利分类具有更优的分类效果。

关键词：专利自动分类 深度学习 词嵌入 专利文本挖掘

分类号：G254.11

DOI：10.13266/j.issn.0252-3116.2020.10.009

1 引言

随着我国知识产权意识的不断提升和知识产权强国建设计划的不断推进，我国专利数量不断取得突破，自2011年起连续8年发明专利申请量均居世界首位[1]。从具体数量来看，根据我国国家知识产权局的专利统计年报，2016年、2017年和2018年我国国内专利申请量分别为3 305 225件、3 536 333件和4 146 772件[2]。如此大规模的专利申请数量得益于我国科技实力的逐步增强，但也给专利审查、管理和分析挖掘带来了巨大挑战。

专利分类是对海量专利文献组织、检索、分析和管理的有效手段，目前国际上应用广泛的专利分类体系包括国际专利分类IPC（International Patent Classification）、美国专利分类USPC（U.S. Patent Classification）、欧洲专利分类ECLA（European Classification System）、日本专利分类FI/F-term和联合专利分类CPC（cooperative patent classification）等，此外还有一些根据特定需求场景定制的个性化分类，如中科院拍卖专利技术分类、专利国民经济产业分类[3]等。通过这些分类，技术人员能够更快速地定位专利信息，专利分析人员能够高效检索专利数据集，市场运营人员能够精准定位潜在转移转化专利。但是，现阶段专利分类号的划分工作目前基本依靠审查员或领域技术人员划分，在当前专利数量爆炸式增长的背景下，这项工作的工作量和工作压力愈发变大。因此，伴随着智能技术的迅猛发展，引入先进智能技术探索自动化专利分类方式对于减轻专利分类工作量、提升分类工作效率具有重要的现实意义。

作为近年来智能技术的代表性技术——深度学习技术，在文本分类任务上不断被应用并取得更优效果[4-6]。基于此，本研究结合专利文本相对格式化的撰写和行文特点，研究基于深度学习技术的专利自动分类方法并进行分类效果评价。

2 专利自动分类研究现状

专利自动分类是计算机基于特定规则、元数据或文本内容等特征自动地为专利分配一个或几个专利分

* 本文系中国科学院青年人才项目"基于深度学习的专利所属产业分类"（项目编号：G180161001）研究成果之一。

作者简介：吕璐成（ORCID：0000-0002-2318-1073），助理研究员，博士研究生，E-mail：lucheng918@126.com；韩涛（ORCID：0000-0001-5955-7813），研究员，博士，硕士生导师；周健（ORCID：0000-0001-8674-6062），博士研究生；赵亚娟（ORCID：0000-0003-3501-8131），研究员，博士，博士生导师。

收稿日期：2019-11-11 修回日期：2019-12-27 本文起止页码：75-85 本文责任编辑：杜杏叶

类号的过程。

专利自动分类研究从分类体系角度可分为基于现有专利分类体系进行分类以及基于个性化分类体系分类两种。基于现有专利分类体系进行分类的研究主要围绕 IPC[7-9]、USPC[10]、ECLA[11-12]、FI/F-term[13-14] 等国际通用分类体系为依据进行分类展开;基于个性化分类体系进行分类的研究主要围绕基于 TRIZ 等经典理论体系或根据特定需求定制的分类体系作为分类依据进行分类,如 C. HE[15-16]、胡正银[17]、翟继强[18]开展了基于面向 TRIZ 设计的分类体系的专利分类,刘龙繁等[19]基于面向产品创新设计的专利功能基分类体系开展自动分类研究,X. ZHANG[20]基于电动汽车领域分类体系(专家划分)开展自动分类研究。

专利自动分类研究从分类方法角度可分为基于特定规则、基于引证关系和基于文本内容挖掘的分类方法三类。基于特定规则分类方面,如 C. HE[16]基于关联规则挖掘方法识别类目规则,进而构建自动分类器;基于引证关系方面,如 S. CHANG 等[21]基于专利引证关系对专利进行聚类并对类簇涉及技术进行解读进而构建分类体系,K. LAI 等[22]基于基础专利的共被引关系采用因子分析的方法建立分类体系;基于文本内容挖掘分类的研究数量较多且持续受到关注,以下进行详细论述。

基于专利文本内容挖掘进行自动分类属于自然语言处理(NLP, Natural Language Processing)中的文本分类任务,该过程的经典方法是采用机器学习方法,通过特征工程的手段,确定专利分类潜在依据特征,进而采用贝叶斯分类器、SVM、逻辑回归等机器学习算法进行自动分类。此类方法常用的特征是词袋特征,即采用词袋模型(Bag of Words)将专利文本表示为所包含词汇的词频向量[8, 23],但由于单纯词频表示带来的无效词(如虚词、连词等功能词)高频噪声问题,后来采用词频逆文档频率(TFIDF)取代原始向量中的词频的方法被广泛应用,如贾杉杉等[24]使用从专利申请书中提取的 TFIDF 特征,分别训练朴素贝叶斯、支持向量机、AdaBoost 分类器预测 IPC 分类号。此外,一些新的特征也在被不断引入进而提升分类效果,如 S. VERBERNE 等[25]在专利特征词的基础上加上特征词语义三元组信息进而改善分类准确率;J. STUTZKI 等[26]引入专利申请人地理位置地理数据特征,使用 KNN 和采用一对其余(one-versus-rest)策略的 SVM 分类器进行专利分类;S. LIM 等[27]同时在标题、摘要、权利要求、技术领域和背景技术信息中抽取特征进而改善专利文本分类效果。基于经典机器学习方法的专利自动分类依赖研究人员手工构建特征来取得更好的分类效果。但是,以词袋模型为代表的特征表示方式丢失了专利文本中词义信息、语序信息等语义信息,例如两篇同类别的文档可能由于用词描述方式不同而无法准确分类。

近年来,随着深度学习技术的崛起和在专利情报研究中的不断应用,在基于专利文本内容挖掘的专利自动分类这一研究场景中,也产生了一系列研究成果。如马双刚[28]基于深度学习理论设计了降噪自动编码器(DAE)和 SVM 算法结合的自动分类方法,并选取计算机领域的六个 IPC 类别进行分类效果验证;胡杰等[29]提出了一种基于卷积神经网络与随机森林算法的专利文本分类模型,应用于英文机械专利文本分类场景;马建红等[30]构建基于 attention 的双向 LSTM(Long Short-Term Memory,长短期记忆网络)模型,对以 100 个专利应用效应作为类标签的机械物理类专利文本进行模型训练和分类测试;S. B. Li 等[31]提出一种基于卷积神经网络和 word embedding 技术的 DeepPatent 方法对英文专利的 IPC 小类(Sub-class)分类号进行自动分类;肖立中等[32]采用 Word2Vec 模型和 LSTM 模型发明了一种安全领域中文专利文本的分类方法,该方法在安全领域中文专利测试集的准确率得到较大提升。综上可知,国内外基于深度学习技术开展专利分类的方法改进与应用研究已经取得一些成果,但是这些研究基本是围绕改进后的特定深度学习方法与传统机器学习方法的分类效果进行比对的,进行的尚未形成有层次的方法优化逻辑体系。

因此,本文面向当前国内专利审查和专利情报分析工作中对于大规模国内专利文献分类的客观需求,针对传统方法存在的缺陷,考虑专利文本语序特征、上下文特征以及分类关键特征,引入深度学习技术,有层次、成体系地设计了 7 种专利深度学习分类方法,并以中国专利为例,选取 IPC 主分类号的部(Section)作为分类依据,比较了 10 种自动分类方法在中文专利分类任务中的表现,从而分析研判深度学习技术用于专利自动分类的效果,为专利分类工作助力。

3 方法设计

3.1 相关概念辨析

文本向量表示和分类模型是开展文本分类的基础,以下对本研究选取的相关文本向量表示方法和基础分类模型的概念进行阐释。

3.1.1 文本向量表示

文本向量表示的经典做法是采用向量空间模型

(VSM),即将文本表示成实数值分量所构成的向量,分量可以采用词频或者词的TFIDF值表示。由于词频无法表示词的重要程度,而TFIDF可以用于评估一个词对语料库中一份文件的重要程度,因此目前以TF-IDF进行向量表示的做法较为广泛。虽然向量空间模型具有清晰明确易解释的优点,但是其存在向量维度随着词表增大而增大且向量高度稀疏的问题,同时其也无法处理同义词、近义词的语义问题,例如“术语抽取”“语义挖掘”和“太阳能电池”三个词在TFIDF特征向量中代表三个特征维度,这些词特征之间虽然有一定语义相似关系但在TFIDF中却无法度量。对此,Google公司在2013年推出的Word2Vec技术能够使用低维度连续分布式向量来表示一个词的语义,并且能够有效表征同义词、近义词等语义相近的词之间的相似关系,因此在文本向量表示方面具有更高的可用性。

本研究采用基于Word2Vec词向量的专利文本向量表示方法进行深度学习模型的专利文本向量表示,并采用基于TFIDF的专利文本向量表示方法作为对照模型的文本向量表示方法。

3.1.2 基础模型

(1)ANN模型。ANN模型是神经网络中基础的全连接层模型,模型包括三层,分别是输入层,隐藏层和输出层,层和层之间是全连接。ANN模型能够将连续分布式向量表示映射到专利文本的标记空间,本质也是能够对专利文本的词向量表示作进一步的特征变换,突出融合相关特征。

(2)TextCNN模型。TextCNN是将卷积神经网络模型用于NLP任务的代表性模型[33],其将卷积神经网络和语言模型的N-gram思想结合起来,通过不同大小的卷积核对文本向量进行不同维度的上下文特征提取,然后通过最大池化操作对提取出的文本向量进行特征强化操作,从而提升文本特征提取能力,提升文本分类效果。

假设一段文本词向量表示 $X=\{x_1,x_2,\cdots,x_m\}$,$x_i\in R^d$,TextCNN分成三个阶段,卷积层,池化层和全连接层,见图1。输入层是 x_i,代表某件专利文本的词向量。

$$x_{1:m}=x_1\oplus x_2\oplus\cdots\oplus x_m \qquad 公式(1)$$

$\oplus$代表拼接操作,$x_{i:j}$代表专利文本中的第i到j个词向量的拼接。将 $x_{1:m}$ 作为卷积层的输入。卷积层结合N-gram思想采用尺寸分别为2*d,3*d,4*d和5*d四种大小的卷积核对 $x_{1:m}$ 进行不同维度的局部特征抽取,公式如下:

$$c_i=f(w*x_{i:i+h-1}+b) \qquad 公式(2)$$

$$C=[c_1,c_2,\cdots,c_{m-h+1}] \qquad 公式(3)$$

w 为卷积核的参数,h 为卷积核的高度,$w\in R^{h*d}$,b 为偏置,$b\in R$,$f(*)$ 为Relu激活函数,C 为卷积层的一个输出,$C\in R^{m-h+1}$。

接着采用最大池化层强化特征,即 $\hat{C}=\max(C)$,最后将池化层结果并拼接起来经过全连接层得到TextCNN的输出,也可直接通过池化层的输出直接进行Softmax分类操作。

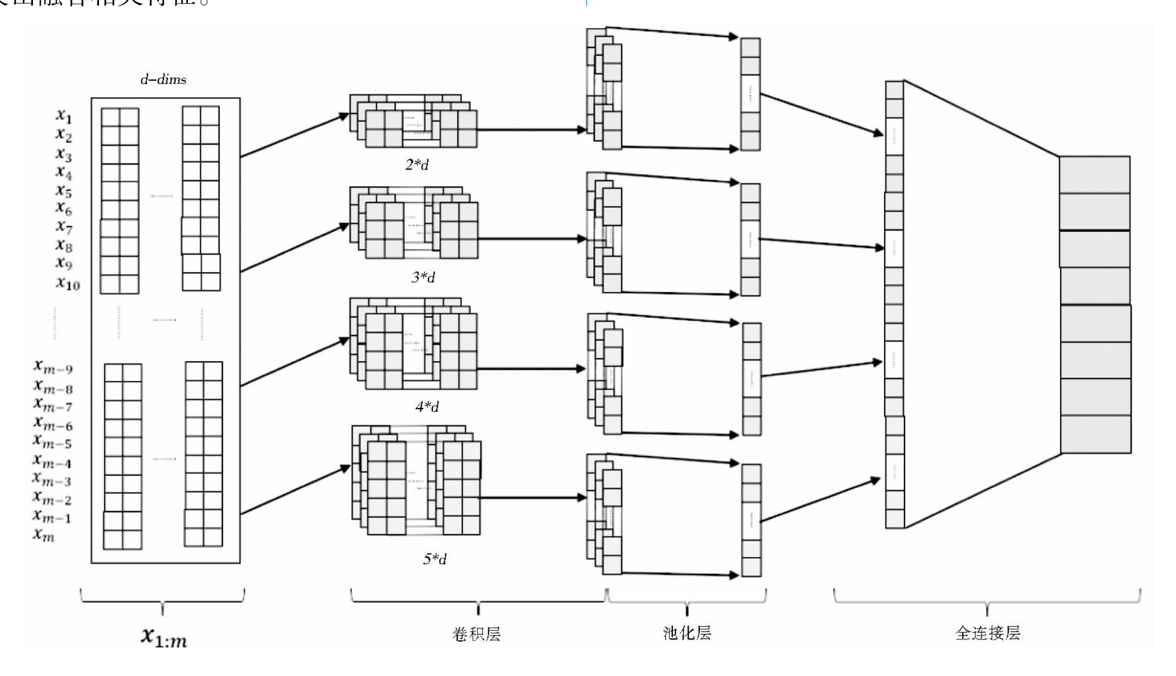


图1 TextCNN结构

（3）GRU/BiGRU 模型。GRU 是循环神经网络的一种变体，和 LSTM 类似是一种特殊的循环神经网络结构[34-35]。标准的循环神经网络单元中，只包含一个 tanh 层进行重复学习，所以会出现梯度消失或者梯度爆炸的问题。为了解决这些问题，基于门控的循环神经网络例如 LSTM 和 GRU 就被提出。本文选择 RNN 模型中的 GRU 而非 LSTM 的原因是实验表明 GRU 和 LSTM 的效果相差不大，且 GRU 有更少的训练参数，由 LSTM 的三个门限单元变成了 GRU 的两个门限单元，因此相对容易训练，并且过拟合的问题相对较少。

GRU 通过共享参数的 GRU 单元依次对文本向量进行计算，并通过最后一步的隐藏向量作为原文本向量的序列化表示，进行分类操作。GRU 单元中包含两个门：更新门和复位门，见图 2。GRU 单元的计算公式如下所示：

$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z) \quad \text{公式(4)}$$

$$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r) \quad \text{公式(5)}$$

$$\hat{h}_t = \tanh(W_h x_t + U_h(r_t h_{t-1}) + b_h) \quad \text{公式(6)}$$

$$h_t = (1 - z_t) h_{t-1} + z_t \hat{h}_t \quad \text{公式(7)}$$

其中 h_{t-1}是 $t-1$ 时刻 GRU 单元的输出，x_t 是 t 时刻 GRU 单元的输入，z_t 是更新门的输出，W_z，U_z 和 b_z 是更新门的权重，r_t 是复位门的输出，W_r，U_r 和 b_r 是复位门的权重，W_h，U_h 和 b_h 是输出门的权重，h_t 是 t 时刻 GRU 单元的输出。

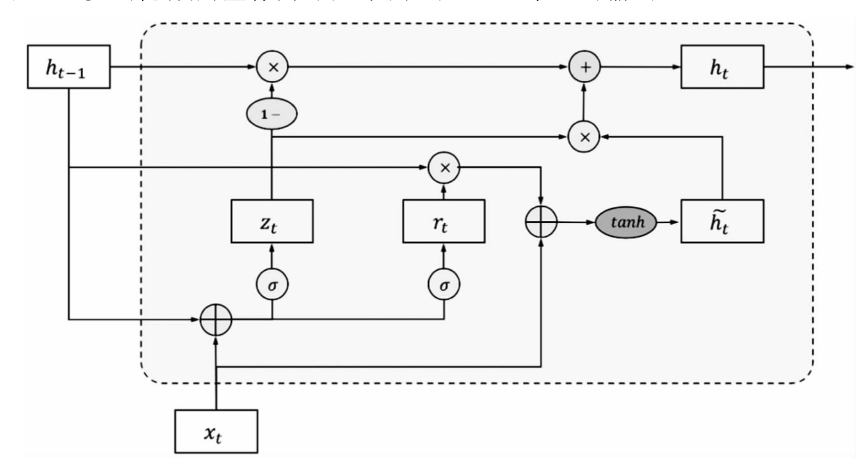


图 2 GRU cell 内部结构

假设一段文本词向量表示 $X = \{x_1, x_2, \cdots, x_m\}$，经过 GRU 层编码得到向量文本的隐藏表示为 $H = \{h_1, h_2, \cdots, h_m\}$，其中：

$$h_t = GRU(h_{t-1}, x_t), t \in [1, m] \quad \text{公式(8)}$$

h_t 代表第 t 步的隐藏表示。

GRU 在对句子进行编码的时候考虑了词的上文信息，但是往往有的时候一个词下文的词对词的编码也会起到一定作用，所以这个时候就考虑到了双向循环神经网络对句子进行编码。BiGRU 在 GRU 的基础上通过 GRU 单元分别对句子进行正向和反向编码，正向反向编码的 GRU 单元内部的参数不共享，文本向量经过 BiGRU 层编码得到文本的隐藏向量表示为 $H = \overrightarrow{H} \oplus \overleftarrow{H} = \{\overrightarrow{h_1} \oplus \overleftarrow{h_1}, \overrightarrow{h_2} \oplus \overleftarrow{h_2}, \cdots, \overrightarrow{h_m} \oplus \overleftarrow{h_m}\}$，其中：

$$\overrightarrow{h_t} = \overrightarrow{GRU(x_t)}, t \in [1, m] \quad \text{公式(9)}$$

$$\overleftarrow{h_t} = \overleftarrow{GRU(x_t)}, t \in [m, 1] \quad \text{公式(10)}$$

$\overrightarrow{h_i}$和$\overleftarrow{h_i}$总结了 x_i 的上下文信息，但是注意力还是聚焦在 x_i 上。

语序信息是文本的重要特征，由于 GRU 循环神经网络能够对文本向量进行序列化建模并表征语序信息，因此将其引入到专利文本分类中。又因 GRU 仅考虑文本向量上文语序特征，而 BiGRU 考虑上下文语序特征，为了研究分类效果两者在本研究中均有考虑。

（4）Attention 机制。Attention 机制源于视觉图像领域，后应用到 NLP 领域并不断取得新进展[36]。现经不断改进已形成多种变体，但核心思想基本为通过给向量分配不同的权重系数来突出对结果影响较大的特征。

本文采用的 Attention 方法基本思路为：假设原文本词向量表示 $X = \{x_1, x_2, \cdots, x_m\}$，经过循环神经网络得到每一步的隐藏表示 $H = \{h_1, h_2, \cdots, h_m\}$。通过给循环神经网络得到的每一步的隐藏表示赋一个权重向量 a_t，公式如下：

$$u_t = \tanh(W_a h_t + b_a) \quad \text{公式(11)}$$

$$a_t = \frac{\exp(u_t^T U)}{\sum_t \exp(u_t^T U)} \quad \text{公式(12)}$$

$$c_t = a_t h_t \quad \text{公式(13)}$$

u_t 是 h_t 的隐藏表示，$U = \{u_1, u_2, \cdots, u_m\}$，$a_t$ 是通

过隐藏表示 u_t 计算得到的归一化后的概率权重，通过概率权重和原始循环神经网络的隐藏表示得到基于权重的文本向量 $C=\{c_1,c_2,\cdots,c_m\}$。

Attention 机制通过自学习到的一个权重矩阵对文本的词向量或其他隐藏表示向量不同位置赋予不同权重，旨在突出关键特征，忽略无用特征。让模型更加注重于那些对结果影响大的部分。Attention 机制不受句子长度的限制，能够突出长句中的关键特征，因此本文在分类模型中引入了 Attention 机制。

3.2 分类模型设计

基于上述文本向量表示方法和基础模型，本研究结合专利文本相对结构化的撰写特点，共设计了 7 种深度学习模型，并设计 3 种经典机器学习模型作为对照，以判别深度学习模型的分类效果，如表 1 所示：

表 1　本研究设计的 10 种专利自动分类模型

模型类型	模型	模型特点
经典机器学习模型	TFIDF + LR	基线模型
	TFIDF + DT	
	TFIDF + RF	
深度学习模型	Word2Vec + ANN	解决近义词、同义词特征问题
	Word2Vec + TextCNN	强化上下文特征
	Word2Vec + GRU	考虑语序特征
	Word2Vec + BiGRU	考虑双向语序特征，解决一词多义问题
	Word2Vec + BiGRU + TextCNN	同时考虑上下文特征和双向语序特征
	Word2Vec + ATT	强化关键特征
	Word2Vec + BiGRU + ATT + TextCNN	同时考虑上下文特征和双向语序特征，强化关键特征

3.2.1 TFIDF + 经典机器学习模型

“TFIDF + 经典机器学习模型”是本研究设计的基线对照模型，其以专利文本的 TFIDF 特征向量作为输入，采用逻辑回归模型(LR)、决策树模型(DT)和随机森林模型(RF)三种经典分类模型来训练专利文本自动分类器。

3.2.2 Word2Vec + ANN

“Word2Vec + ANN”是为了区分专利文本中近义词同义词以获得更好的自动分类效果的一种分类模型。

假设一篇专利文本的词向量表示 $X=\{x_1,x_2,\cdots,x_m\}$，$x_i\in R^d$，Word2Vec + ANN 模型计算公式简单描述如下：

$$X=Flatten(X) \qquad 公式(14)$$

$$H=tanh(W_h\ X+b_h) \qquad 公式(15)$$

$$O=softmax(W_o\ H+b_o) \qquad 公式(16)$$

$$\hat{y}=argmax(O) \qquad 公式(17)$$

$Flatten(*)$ 为向量展开操作，将高维向量展开成一维向量，H 代表隐藏层输出，O 代表输出层输出，$\hat{y}$ 代表模型预测出的标签，W 和 b 为网络权重参数。

3.2.3 Word2Vec + TextCNN

Word2Vec + ANN 模型直接将专利文本向量展开成一维，这个过程会丢掉文本上下文、语序等很多语义信息，而且无法发挥深度学习的特征提取和表示能力。为了提取和强化局部上下文的特征，本文提出了“Word2Vec + TextCNN”模型。

假设一篇专利文本的词向量表示 $X=\{x_1,x_2,\cdots,x_m\}$，$x_i\in R^d$，Word2Vec + TextCNN 模型计算公式简单描述如下：

$$C_j=Conv1d(X,j),j\in[2,5] \qquad 公式(18)$$

$$P_j=Maxpooling(C_j),j\in[2,5] \qquad 公式(19)$$

$$O_{conv}=P_2\oplus P_3\oplus P_4\oplus P_5 \qquad 公式(20)$$

$$O=softmax(W_o\ O_{conv}+b_o) \qquad 公式(21)$$

$$\hat{y}=argmax(O) \qquad 公式(22)$$

$Conv1d(X,j)$ 中，X 代表输入 TextCNN 的专利文本词向量，j 代表卷积核的大小，$Maxpooling(*)$ 代表最大池化操作，$\oplus$ 代表向量拼接操作，W_o 和 b_o 代表输出层的网络参数，$\hat{y}$ 代表模型预测的类别。

3.2.4 Word2Vec + GRU

Word2Vec + TextCNN 模型提取并强化了当前词和邻近词的特征，但是没有考虑专利文本全局的语序特征。对于 NLP 任务，语序特征是一项很独特的特征。对于图像来说，调换某两个位置的像素值可能对结果不会产生特别大的影响，但对于文本来说，调换某两个词的顺序可能会使得句子的语义产生很大变化。所以针对句子语序建模的问题，本文提出了“Word2Vec + GRU”模型。

假设一篇专利文本的词向量表示 $X=\{x_1,x_2,\cdots,x_m\}$，$x_i\in R^d$，Word2Vec + GRU 模型计算公式简单描述如下：

$$h_t = GRU(h_{t-1}, x_t), t \in [1, m]$$ 公式(23)

$$O = softmax(W_o h_m + b_o)$$ 公式(24)

$$\hat{y} = argmax(O)$$ 公式(25)

$GRU(h_{t-1}, x_t)$中，h_{t-1}代表 $t-1$ 步的隐藏表示，x_t 代表当前输入，h_m 代表最后一步的隐藏表示，O 代表输出层的输出，W_o 和 b_o 代表输出层的网络参数，$\hat{y}$ 代表模型预测的类别。

3.2.5 Word2Vec + BiGRU

Word2Vec + GRU 模型考虑了正向的语序特征，也就是在第 t 步时间做计算的时候只会考虑到前 $t-1$ 步的历史状态，而不会考虑从 $t+1$ 之后的信息，所以使用 GRU 对专利文本的词向量做序列建模可能是不全面的；而 BiGRU 对专利文本的词向量做序列建模时同时进行双向建模，不仅考虑了双向语序特征，而且考虑了前后文语义特征。基于此，本文提出了"Word2Vec + BiGRU"模型。

假设一篇专利文本的词向量表示 $X = \{x_1, x_2, \cdots, x_m\}, x_i \in R^d$，Word2Vec + BiGRU 模型计算公式描述如下：

$$\overrightarrow{h_t} = \overrightarrow{GRU}(h_{t-1}, \overrightarrow{x_t}), t \in [1, m]$$ 公式(26)

$$\overleftarrow{h_t} = \overleftarrow{GRU}(h_{t+1}, \overleftarrow{x_t}), t \in [m, 1]$$ 公式(27)

$$O = softmax(W_o(\overrightarrow{h_m} \oplus \overleftarrow{h_1}) + b_o)$$ 公式(28)

$$\hat{y} = argmax(O)$$ 公式(29)

$\overrightarrow{h_t}$为正向建模第 t 步的隐藏表示，$\overleftarrow{h_t}$为逆向建模第 t 步的隐藏表示，根据上述公式可知，正向建模的最后一步隐藏表示为$\overrightarrow{h_m}$，逆向建模的最后一步隐藏表示为$\overleftarrow{h_1}$，所以$\overrightarrow{h_m} \oplus \overleftarrow{h_1}$为 BiGRU 层的隐藏输出，$O$ 代表输出层的输出，W_o 和 b_o 代表输出层的网络参数，$\hat{y}$ 代表模型预测的类别。

3.2.6 Word2Vec + BiGRU + TextCNN

Word2Vec + BiGRU 模型较为完善的考虑了双向语序特征，同时根据上下文的语义信息动态调整了词向量表示，在一定程度上解决了一词多义的问题。但是没有提取和强化当前词的上下特征，这会使得一些隐藏的关键特征没有明显的突出出来，导致分类结果不理想。所以本文结合提取序列特征的 BiGRU 模型和强化上下文特征的 TextCNN 模型提出了"Word2Vec + BiGRU + TextCNN"模型。该模型首先使用 BiGRU 对专利文本的向量表示进行双向建模，得到根据上下文动态调整词向量后的隐藏表示，然后以该隐藏表示作为 TextCNN 的输入，通过卷积神经网络提取特征和池化层强化特征。

假设一篇专利文本的词向量表示 $X = \{x_1, x_2, \cdots, x_m\}, x_i \in R^d$，Word2Vec + BiGRU + TextCNN 模型计算公式简单描述如下：

$$\overrightarrow{h_t} = \overrightarrow{GRU}(h_{t-1}, \overrightarrow{x_t}), t \in [1, m]$$ 公式(30)

$$\overleftarrow{h_t} = \overleftarrow{GRU}(h_{t+1}, \overleftarrow{x_t}), t \in [m, 1]$$ 公式(31)

$$H = \overrightarrow{H} \oplus \overleftarrow{H} = \{\overrightarrow{h_1} \oplus \overleftarrow{h_1}, \overrightarrow{h_2} \oplus \overleftarrow{h_2}, \cdots, \overrightarrow{h_m} \oplus \overleftarrow{h_m}\}$$ 公式(32)

$$C_j = Conv1d(H, j), j \in [2, 5]$$ 公式(33)

$$P_j = Maxpooling(C_j), j \in [2, 5]$$ 公式(34)

$$O_{conv} = P_2 \oplus P_3 \oplus P_4 \oplus P_5$$ 公式(35)

$$O = softmax(W_o O_{conv} + b_o)$$ 公式(36)

$$\hat{y} = argmax(O)$$ 公式(37)

其中 H 代表 BiGRU 层的隐藏表示，由正向隐藏表示和逆向隐藏表示构成。

3.2.7 Word2Vec + Attention

TextCNN 能够捕捉的局部的上下文关键特征，BiGRU 对序列特征进行建模和提取，但是这两种方法有一定局限，那就是对于长距离的关键特征无法有效地捕捉并强化。由于 Attention 机制能够突出长句中的关键特征，因此本文提出了"Word2Vec + Attention"模型，即通过词向量训练得到一组对应于词向量的特征权重矩阵，通过基于权重的词向量加权得到最后的文本向量表示。

假设一篇专利文本的词向量表示 $X = \{x_1, x_2, \cdots, x_m\}, x_i \in R^d$，Word2Vec + Attention 模型计算公式简单描述如下：

$$u_t = \tanh(W_a x_t + b_a)$$ 公式(38)

$$a_t = \frac{\exp(u_t^T U)}{\sum_t \exp(u_t^T U)}$$ 公式(39)

$$c = \sum_t a_t x_t$$ 公式(40)

$$O = softmax(W_o c + b_o)$$ 公式(41)

$$\hat{y} = argmax(O)$$ 公式(42)

u_t 是由 x_t 计算得到的隐藏表示，a_t 则是由隐藏表示归一化得到的权重向量，W 和 b 为网络参数，c 代表根据 Attention 权重矩阵加权得到的文本向量表示。

3.2.8 Word2Vec + BiGRU + Attention + TextCNN

综合以上六种深度学习模型的特点，融合能够对专利文本向量进行双向序列化建模的 BiGRU 模型、使用卷积神经网络提取局部特征通过池化层强化特征的 TextCNN 模型、能够忽略距离强化关键特征的 Attention 机制，提出本文第七种深度学习模型——"Word2Vec + BiGRU + Attention + TextCNN"模型。首先通过 BiGRU 对词向量进行动态调整，然后使用 Attention 机制对

BiGRU 输出的隐藏表示进行权重调整,最后将调整后的隐藏表示作为 TextCNN 的输入。

假设一篇专利文本的词向量表示 $X=\{x_1,x_2,\cdots,x_m\}$,$x_i\in R^d$,Word2Vec + BiGRU + Attention + TextCNN 模型计算公式简单描述如下:

$$\overrightarrow{h_t}=\overrightarrow{GRU}(h_{t-1},\overrightarrow{x_t}),t\in[1,m] \quad 公式(43)$$

$$\overleftarrow{h_t}=\overleftarrow{GRU}(h_{t+1},\overleftarrow{x_t}),t\in[m,1] \quad 公式(44)$$

$$H=\overrightarrow{H}\oplus\overleftarrow{H}=\{\overrightarrow{h_1}\oplus\overleftarrow{h_1},\overrightarrow{h_2}\oplus\overleftarrow{h_2},\cdots,\overrightarrow{h_m}\oplus\overleftarrow{h_m}\} \quad 公式(45)$$

$$u_t=\tanh(W_a(\overrightarrow{h_t}\oplus\overleftarrow{h_t})+b_a) \quad 公式(46)$$

$$a_t=\frac{\exp(u_t^T U)}{\sum_t \exp(u_t^T U)} \quad 公式(47)$$

$$c_t=a_t(\overrightarrow{h_t}\oplus\overleftarrow{h_t}) \quad 公式(48)$$

$$C=\{c_1,c_2,\cdots,c_t\} \quad 公式(49)$$

$$Conv_j=Conv1d(C,j),j\in[2,5] \quad 公式(50)$$

$$P_j=Maxpooling(Conv_j),j\in[2,5] \quad 公式(51)$$

$$O_{conv}=P_2\oplus P_3\oplus P_4\oplus P_5 \quad 公式(52)$$

$$O=softmax(W_o O_{conv}+b_o) \quad 公式(53)$$

$$\hat{y}=argmax(O) \quad 公式(54)$$

其中 H 是 BiGRU 层输出的隐藏表示,C 是 Attention 层输出的隐藏表示,不同于 Word2Vec + Attention 模型直接加权处理。

3.3 模型效果评估指标

本文选用三种评价指标来评估模型效果,分别是准确率,召回率和 F1 值,采用宏平均指标计算。宏平均指标是先对每一个类别计算统计指标,然后对所有类别计算算数平均值,公式如下。

$$p_k=\frac{预测出的\ k\ 类别并且正确的样本数}{预测出的\ k\ 类别的样本数} \quad 公式(55)$$

$$r_k=\frac{预测出的\ k\ 类别并且正确的样本数}{测试样本中的\ k\ 类别的数目} \quad 公式(56)$$

$$F1_k=\frac{2*p_k*r_k}{p_k+r_k} \quad 公式(57)$$

$$P=\frac{1}{K}\sum_{k=1}^{K}p_k \quad 公式(58)$$

$$R=\frac{1}{K}\sum_{k=1}^{K}r_k \quad 公式(59)$$

$$F1=\frac{1}{K}\sum_{k=1}^{K}F1_k \quad 公式(60)$$

其中 K 代表总类别数目,p_k,r_k 和 $F1_k$ 分别代表 k 类别的准确率,召回率和 F1 值。准确率 p_k 是衡量正确划分到 k 类别的文本占划分到 k 类别的文本的比例,p_k 越大说明模型对于 k 类别样本分类越准确。召回率 r_k 是衡量正确划分到 k 类别的文本占实际文本中 k 类别的文本的比例,r_k 越大说明模型在 k 类别上漏掉的样本越少。$F1_k$ 综合考虑准确率和召回率,值越高说明 k 类别的分类效果越理想。

3.4 方法实施流程

本文设计的专利自动分类方法流程分为 5 个步骤,分别是:数据集构建,文本预处理,文本向量化,模型训练及调参和模型分类效果评估,如图 3 所示:

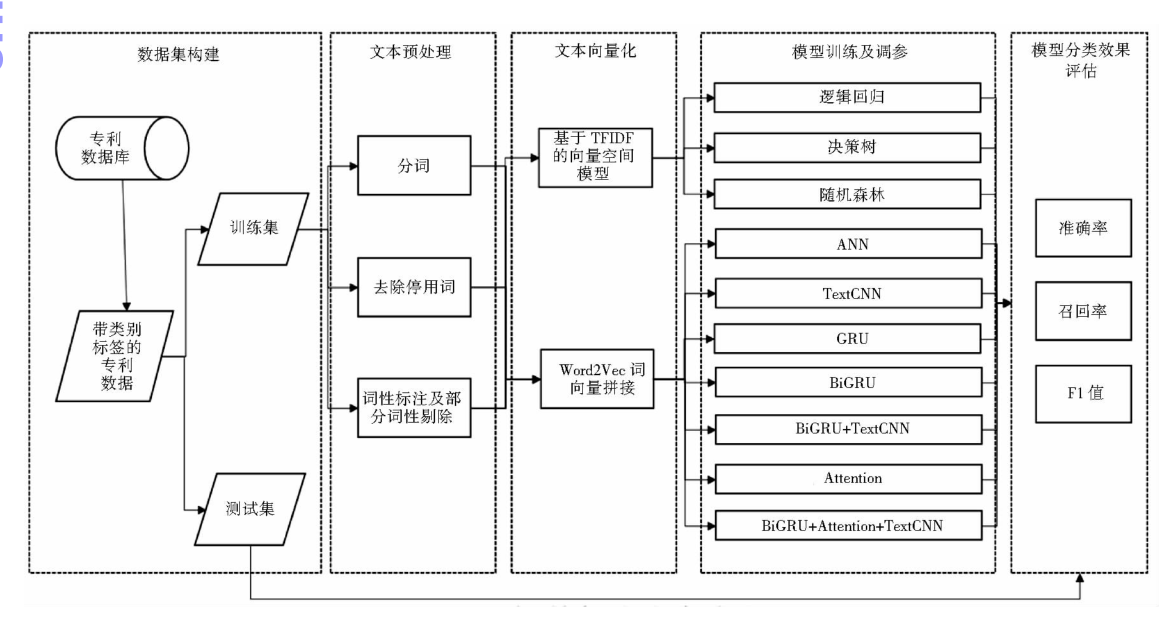


图 3 专利自动分类流程

图 3 中各环节具体释义如下：

（1）数据集构建。从专利数据库中提取出适当数量的带类别标签的专利，作为原始数据集集合，将原始数据集集合划分成训练集和测试集两个部分：训练集和测试集。训练集是用于训练专类自动分类模型的数据集，为了更好的训练模型，训练集中部分比例会划分出来作为验证集配合训练模型。测试集合是用于评价已训练好的专利自动分类模型。

（2）文本预处理。包括分词、去除停用词和词性标注去特定词性三个步骤。对中文而言，字是最小的字符单元，而最小的语义单元是词，所以为了模型能够从语义的角度处理文本，取得更好的效果，在文本预处理阶段第一步先对数据集中的专利文本部分分词处理。专利文本的分词结果中会存在一些噪声词，如特殊字符或者无实际意义的虚词，这些词通过去除停用词和词性标注去除特定词性（仅保留名词、动词、形容词等实词）的方式剔除。

（3）文本向量化。采用基于 TF-IDF 的向量空间模型进行三种传统基线模型的文本向量化，采用基于 Word2Vec 词向量的向量拼接法进行七种深度学习模型的专利文本向量化。

（4）模型训练及调参。采用前文所述的 10 种专利自动分类模型进行模型训练和参数调优，训练过程保留在验证集上效果最好的模型。

（5）模型分类效果评估。将 10 种分类模型在测试集上进行测试，评估其在准确率、召回率、F1 值指标上的表现。

4 实验结果及效果分析

4.1 分类依据及实验数据

本文选取专利 IPC 主分类号的“部”作为分类依据（各部含义见表 2）。从中科院专利在线分析系统随机抽取 80 000 条专利数据作为数据集，将数据集划分成三个部分：50 000 条作为训练集，10 000 条作为验证集，20 000 条作为测试集。采用的 Word2Vec 词向量模型基于 CBOW 模型从三千多万中文专利数据训练获得，训练参数 size 为 300，min_count 为 40，window 为 10，sample 为 1e－3。

表 2 分类依据

IPC 部	技术含义
A	人类生活必需
B	作业；运输
C	化学；冶金
D	纺织；造纸
E	固定建筑物
F	机械工程；照明；加热；武器；爆破
G	物理
H	电学

4.2 分类结果及结果分析

本文采用 Mini-batch 训练，经过试验分析，最终选择每个 Mini-batch 样本大小为 200，词向量维度是 300，循环神经网络输出维度是 300，卷积神经网络输出维度是 300，卷积核大小分别为 2，3，4，5。模型迭代至验证集上结果收敛为止，并且保留在验证集上效果最好的模型结果。结果如表 3 所示：

表 3 模型自动分类结果

模型类型	模型	准确率	召回率	F1
经典机器学习模型	TFIDF + LR	0.780 5	0.778 4	0.778 6
	TFIDF + DT	0.575 9	0.574 0	0.574 8
	TFIDF + RF	0.715 6	0.711 7	0.708 2
深度学习模型	Word2Vec + ANN	0.730 0	0.730 1	0.730 0
	Word2Vec + TextCNN	0.810 3	0.807 5	0.807 5
	Word2Vec + GRU	0.808 3	0.809 1	0.808 1
	Word2Vec + BiGRU	0.812 0	0.811 7	0.811 4
	Word2Vec + BiGRU + TextCNN	0.822 0	0.818 3	0.817 5
	Word2Vec + ATT	0.763 6	0.762 6	0.762 2
	Word2Vec + BiGRU + ATT + TextCNN	0.823 0	0.824 3	0.823 1

按照模型准确率进行排序，得到图 4 所示的 10 种模型的分类准确率、召回率和 F1 值对比结果。

以下对实验结果进行分析阐释：

（1）深度学习模型的效果基本优于经典机器学习

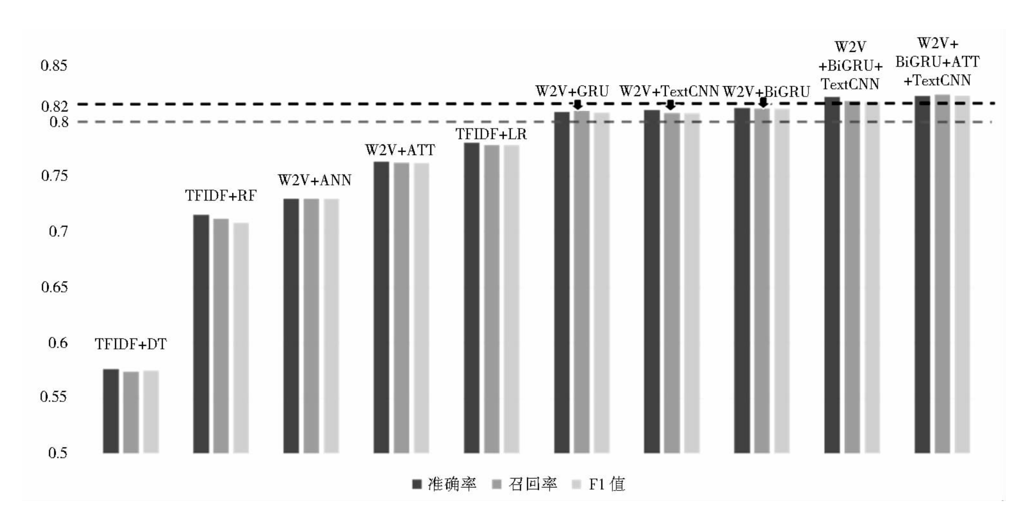


图4　十种模型的分类准确率、召回率和 F1 值对比

模型。除 Word2Vec + ANN 和 Word2Vec + ATT 外的深度学习模型的准确率、召回率和 F1 值均高于 0.8，而三种经典机器学习模型的指标均低于 0.8。由于 ANN 特征表示能力相对较弱、直接在词向量的基础上引入 Attention 机制不能较好地表示隐藏特征，Word2Vec + ANN 和 Word2Vec + ATT 的效果在所有深度学习模型中表现最低，其效果仍旧显著优于 TFIDF + DT 和 TFIDF + RF，这表明通过对文本向量做特征提取和强化对分类结果的优化具有一定促进作用。

（2）考虑上下文特征和语序特征对于分类效果提升有积极作用。TextCNN 模型基于卷积神经网络对专利文本进行上下文的特征抽取和强化；GRU 模型对专利文本进行正向序列建模，强化了上文的序列特征；BiGRU 模型对专利文本进行双向建模，强化了上下文的序列特征。这些特征的考虑使得 Word2Vec + TextCNN、Word2Vec + GRU 和 Word2Vec + BiGRU 均取得了高于 0.8 的指标得分。在此基础上，将 BiGRU 和 TextCNN 模型进行结合，对双向语序特征建模同时考虑上下文特征取得了优于单纯使用 TextCNN 和 BiGRU 的模型效果。

（3）引入 Attention 机制强化关键特征对于分类结果具有正向影响。在 10 种自动分类模型中，Word2Vec + BiGRU + ATT + TextCNN 模型表现最优，这表明在考虑上下文特征和双向语序特征的同时，引入 Attention 机制强化关键特征能够有效提升专利文本分类的效果。

5　结果讨论及展望

本文针对中文专利多分类的问题，基于 TextCNN、GRU、Attention 等技术，设计了 7 种专利自动分类深度学习模型，并与 3 种传统经典自动分类模型进行效果比对评估，最终发现采用考虑语序特征、上下文特征及强化关键特征的深度学习模型较之传统分类模型进行中文专利分类具有更优的分类效果。其中"Word2Vec + BiGRU + ATT + TextCNN"模型在这 10 个模型表现出了最优效果，具有最高的分类准确率、召回率和 F1 值。在当前国家对专利审查工作提速要求的背景下，该模型在一定程度上将对优化提升现有自动分类方法和工具的效果、提升专利分类工作效率及缩短专利审查周期具有借鉴意义和参考价值。

但是本文的研究工作仍旧有待改进，即专利分类问题属于多标签分类问题，而本研究仅选取专利的主分类号开展了单标签多分类问题的研究；同时，专利 IPC 分类包括部、大类、小类、大组和小组五个层级，更细类别的分类意味着类别数量的大幅提升，对分类模型提出了更高要求，本文的研究仅针对"部"开展自动分类研究，之后需对更细层级的分类模型进行研究。此外，本研究提出的方法还可在国际上针对专利分类发布的评测任务和数据（如 NTCIR 评测比赛的专利分类任务[13]）上应用和验证，从而扩大研究的影响力与应用范围。

深度学习技术的发展方兴未艾，例如 Google 于 2018 年发布的 BERT 预训练语言模型已在 11 个 NLP 任务中刷新了记录，这对进一步优化专利自动分类模型的分类效果提供了可能。下一步工作中，将继续研究基于动态文本表示模型的专利自动分类方法，以期获得更优的分类效果。

参考文献:

[1] 央视新闻. 中国发明专利申请量连续 8 年居世界首位[EB/OL]. [2019-08-02]. http://dy.163.com/v2/article/detail/EGM6VQS60511A3UP.html.

[2] 国家知识产权局. 国内专利申请年度状况[EB/OL]. [2019-08-02]. http://www.cnipa.gov.cn/tjxx/jianbao/year2018/a/a3.html.

[3] 田创,赵亚娟. 一种基于相似度的专利与产业类目映射模型——以《国际专利分类》与《国民经济行业分类》为例[J]. 图书情报工作,2016,60(20):123-131.

[4] LAI S W, XU L H, LIU K, et al. Recurrent convolutional neural networks for text classification[C]//Proceedings of the twenty-ninth AAAI conference on artificial intelligence. Austin: AAAI, 2015:2267-2273.

[5] YANG Z C, YANG D Y, DYER C, et al. Hierarchical attention networks for document classification[C]//Proceedings of the 2016 conference of the North American Chapter of the Association for Computational Linguistics: human language technologies. San Diego:NAACL, 2016: 1480-1489.

[6] ZHANG X, ZHAO J, LECUN Y, et al. Character-level convolutional networks for text classification[C]//Advances in neural information processing Systems. Montreal: Neural information processing systems foundation, 2015:649-657.

[7] CHEN Y L, CHANG Y C. A three-phase method for patent classification[J]. Information processing and management, 2012, 48(6):1017-1030.

[8] FALL C J, TORCSVARI A, BENZINEB K, et al. Automated categorization in the international patent classification[C]//ACM SIGIR forum. Toronto: Association for Computing Machinery, 2003, 37(1): 10-25.

[9] TRAPPEY A J C, HSU F C, TRAPPEY C V, et al. Development of a patent document classification and search platform using a back-propagation network[J]. Expert systems with applications, 2006, 31(4):755-765.

[10] HODREA I B, BOT R I, WANKA G. The Rose-Gurewitz-Fox approach applied for patents classification[J]. European journal of operational research, 2006, 173(3):815-826.

[11] KRIER M, FRANCESCO Z. Automatic categorisation applications at the European patent office[J]. World patent information, 2002, 24(3):187-196.

[12] KOSTER C H A, SEUTTER M, BENEY J. Multi-Classification of patent applications with Winnow[C]//International Andrei Ershov memorial conference on perspectives of system informatics. Berlin: Springer Berlin Heidelberg, 2003:546-555.

[13] IWAYAMA M, FUJII A, KANDO N. Overview of classification subtask at NTCIR-5 patent retrieval task[C]//Proceedings of NTCIR-5 workshop meeting. Tokyo:NTCIR, 2005.

[14] KIM J H, CHOI K S. Patent document categorization based on semantic structural information[J]. Information processing and management, 2007, 43(5):1200-1215.

[15] HE C, LOH H T. Grouping of TRIZ inventive principles to facilitate automatic patent classification[J]. Expert systems with applications, 2008, 34(1):788-795.

[16] HE C, LOH H T. Pattern-oriented associative rule-based patent classification[J]. Expert systems with applications, 2010, 37(3):2395-2404.

[17] 胡正银, 方曙, 文奕,等. 面向 TRIZ 的专利自动分类研究[J]. 现代图书情报技术, 2015, 31(1):66-74.

[18] 翟继强, 王克奇. 依据 TRIZ 发明原理的中文专利自动分类[J]. 哈尔滨理工大学学报, 2013, 18(3).

[19] 刘龙繁,李彦,侯超异,等. 基于功能基的专利信息挖掘与自动分类实验研究[J]. 四川大学学报(工程科学版),2016,48(5):105-113.

[20] ZHANG X Y. Interactive patent classification based on multi-classifier fusion and active learning[J]. Neurocomputing, 2014, 127:200-205.

[21] CHANG S B, LAI K K, CHANG S M. Exploring technology diffusion and classification of business methods: using the Patent Citation Network[J]. Technological forecasting and social change, 2009, 76(1):107-117.

[22] LAI K K, WU S J. Using the Patent Co-Citation approach to establish a new patent classification system[J]. Information processing and management, 2005, 41(2):313-330.

[23] 李程雄,丁月华,文贵华. SVM-KNN 组合改进算法在专利文本分类中的应用[J]. 计算机工程与应用,2006(20):193-195,212.

[24] 贾杉杉,刘畅,孙连英,等. 基于多特征多分类器集成的专利自动分类研究[J]. 数据分析与知识发现,2017,1(8):76-84.

[25] VERBERNE S and D'HONDT E. Patent classification experiments with the linguistic classification system LCS in CLEF-IP 2011[C]//CLEF2011 working notes. Amsterdam: CLEF, 2011.

[26] STUTZKI J, MATTHIAS S. Geodata supported classification of patent applications[C]//Proceedings of the third international ACM SIGMOD workshop on managing and mining enriched geo-spatial data. San Francisco: Association for Computing Machinery, 2016: 1-6.

[27] LIM S, KWON Y J. IPC multi-label classification based on the field functionality of patent documents[C]//SIGIR Forum. Gold Coast: Association for Computing Machinery, 2016: 677-691.

[28] 马双刚. 基于深度学习理论与方法的中文专利文本自动分类研究[D]. 苏州:江苏大学,2016.

[29] 胡杰,李少波,于丽娅,等. 基于卷积神经网络与随机森林算法的专利文本分类模型[J]. 科学技术与工程,2018,18(6):268-272.

[30] 马建红，王瑞杨，姚爽，等. 基于深度学习的专利分类方法[J]. 计算机工程，2018，44(10):215－220.

[31] LI S B, HU J, CUI Y X, et al. DeepPatent: Patent classification with convolutional neural networks and word embedding[J]. Scientometrics, 2018, 117(2):721－744.

[32] 肖立中,王广仲,刘源,等. 安全领域专利文本的分类方法[P]. 中国:109033402A. 2018－12－18.

[33] KIM Y. Convolutional Neural Networks for Sentence Classification [C]//Proceedings of the 2014 conference on empirical methods in natural language processing. Doha:EMNLP, 2014:1746－1751.

[34] CHO K, MERRIENBOER B V, GULCEHRE C, et al. Learning phrase representations using RNN encoder——decoder for statistical machine translation[C]//Proceedings of the 2014 conference on empirical methods in natural language processing. Doha: EMNLP, 2014:1724－1734.

[35] CHUNG J, GULCEHRE C, CHO K H, et al. Empirical evaluation of gated recurrent neural networks on sequence modeling[C]//NIPS 2014 deep learning and representation learning workshop. arXiv:1412.3555. Montreal: NIPS, 2014.

[36] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//arXiv:1706.03762. Long Beach: NIPS, 2017.

作者贡献说明:

吕璐成:负责论文框架设计,论文撰写与修改;

韩涛:负责研究方案设计及优化调整;

周健:开展模型训练及效果评价;

赵亚娟:论文选题与设计,提出修改意见。

Research on the Method of Chinese Patent Automatic Classification Based on Deep Learning

Lyu Lucheng[1,2]　Han Tao[1,2]　Zhou Jian[3]　Zhao Yajuan[1,2]

[1] National Science Library, Chinese Academy of Sciences, Beijing 100190

[2] Department of Library, Information and Archives Management, School of Economics and Management, University of Chinese Academy of Sciences, Beijing 100190

[3] Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190

Abstract: [**Purpose/significance**] In order to meet the needs of classifying massive patent automatically in current patent examination and patent information analysis work, this paper studies a series of patent automatic classification methods based on deep learning and compares the classification effects. This will promote the efficiency and effectiveness of patent classification. [**Method/process**] Aiming at the shortcoming of traditional machine learning methods, 7 deep learning models was designed, including Word2Vec + TextCNN, Word2Vec + GRU, Word2Vec + BiGRU, Word2Vec + BiGRU + TextCNN and so on. These models based on the deep learning technology, such as Word2Vec, CNN, RNN, Attention mechanism and so on and considered the characteristics of patent text word order, context features and other key features in classification. Selecting the ‘Section’ of main International Patent Classification (IPC) was as the class labels, the study classified the Chinese patents by above 7 deep learning models and 3 traditional machine learning methods. And there was a comparison about the effect of classification in different models. [**Result/conclusion**] The empirical research indicated that it reached the better effect of Chinese patent classification by using deep learning methods which considered the characteristics of patent text word order, context features and other key features in classification.

Keywords: patent automatic classification　deep learning　word embedding　patent text mining